The Art of Teaching

Statistics

Telling Statistical Tales

Sivashanmugam

Statistics is one of the generic subjects which is taught almost in every branch of Science. Every student of science irrespective of the field of specialization is supposed to be an expert in statistics. But, many of our 'students' suffer with innumeracy. The consequences of innumeracy can be felt by those who teach statistics in the departments of physics, chemistry and biology. Especially, the students of postgraduate biotechnology, microbiology and bioinformatics do not have the basic knowledge of addition, subtraction and exponentials. Such is the integrity of the examination systems of various Indian Universities which award Bachelor of Science degree to students. Therefore, it is absolutely nonsensical to start the statistics course on the basis of the background knowledge of the students. It would be better for you and your students if you **start with what they have, instead of what they do not have**.

Just go to class. Do not take any book with you. You students would declare that the teacher also do not know statistics if you look at the book and instruct. Have some chat on the current sociopolitical issues for some 10-15 minutes. Then, in a humorous manner ask them - **how much they have in their purse**. Ask each student and write the reply of the students in the blackboard. Be cool and the students will enjoy. Suppose if the number of students in the class is 18, write all the 18 replies and tabulate the replies in a nice way.

1. Mansure - Rs. 80
2. Mekala - Rs. 200
3. Siva- Rs. 500
4. Jaabir - Rs. 400
5. Senthilkumar - Rs. 550
6. Rafi - Rs. 350
7. Sheik Mohamed - Rs. 250
8. Vignesh - Rs. 110
9. Alavudeen - Rs. 125
10. Suja Rachel - Rs. 20
11. Elecy Sheelu - Rs. 180
12. Abeena Mathew- Rs. 350
13. Akila - Rs. 230
14. Priya - Rs. 510
15. Stella Monika - Rs. 220
16. Sherline - Rs. 25
17. Anisha Maria - Rs. 170
18. Vinitha - Rs. 190

Ask them to find the total amount of money they have. Make sure that every student is doing the calculations in their notebook. Side by side perform the calculations in the board. Note the fact that many students won't give correct answer. The total amount is Rs. 4460. Ask them to find the total number of students in the class. It is 18. Now, you tell them to **divide the total amount equally to all** the students. Now they have to divide the total money (4460) by the total number of students (18). Upon equal distribution (division), each student will get Rs. 247.78. Now you tell them what is meant by **mean value**. Mean or average is the sum of all values divided by the total number of observations.

Next, you ask them to find who is having the highest amount of money in the class and who is having the lowest amount. Now, tell the to find out the difference between the highest and the lowest values. Tell them that the difference between the highest value and the lowest value is called as Range.

The highest value in the class is Rs. 550. The lowest value in the class is Rs. 20. The range is, therefore, 550-20, which is equal to 530.

Next, ask the students to find out who are having money more than the mean value. Also, ask them to find out who are having money the less than the mean value. Then, ask them to find how much they differ from the mean value. Now, introduce the concept of deviation. Ask them to sum up all the deviations and find out its mean.

The Deviations:

1. Mansure - 80 - 247.78 = **-167.78**
2. Mekala - 200 - 247.78= **- 47.78**
3. Siva- 500 - 247.78=**252.22**
4. Jaabir - 400 - 247.78=**152.22**
5. Senthilkumar - 550 - 247.78=**302.22**
6. Rafi - 350 - 247.78=**102.22**
7. Sheik Mohamed - 250 - 247.78=**2.22**
8. Vignesh - 110 - 247.78=**-137.78**
9. Alavudeen - 125 - 247.78 =**-122.78**
10. Suja Rachel - 20 - 247.78= **-227.78**
11. Elecy Sheelu - 180 - 247.78= **-67.78**
12. Abeena Mathew- 350 - 247.78=**102.22**
13. Akila - 230 - 247.78=**-17.78**
14. Priya - 510 - 247.78=**262.22**
15. Stella Monika - 220 - 247.78=**-27.78**
16. Sherline - 25 - 247.78=**-222.78**
17. Anisha Maria - 170 - 247.78=**-77.78**
18. Vinitha - 190 - 247.78=**-57.78**

The sum of all the deviations is -0.04. The total number of observations is 18. The mean of all deviations, the mean deviation is the sum of all deviations divided by the total number of observations, which is -0.002.

Now, you tell them about the sign problem. Now you explain them, x is nothing but the root of square of x. Give some examples, 6=root of square of 6= square root of 36. So, -0.002 can be converted to by squaring it and the taking the square root of it.

Now, you ask them to square all the individual deviations. Ask them to find mean of square of all the individual deviations. Ask them to find the **R**oot of the **M**ean of **S**quares of all the individual **D**eviations. RMSD is numerically equal to Standard Deviation. Do not ask them to remember the formula of standard deviation, ask them to remember Standard Deviation is nothing but the RMSD.

So, step by step ask them to make columns in the table:

1. first column is to be the Serial Number
2. Next column is to be the value, bottom of the column is to be the sum of values.
3. Next column is to be the Deviation (D of RMSD)
4. Next column is to be the squares of deviation, column bottom has to be the sum of the Squares of Deviation. (SD of RMSD)
5. Next ask them to calculate Mean of the Squares of Deviation (MSD of RMSD) by dividing the Sum of the Squares of Deviation by the total number of observations.
6. Next ask the to find the Root of Mean of Squares of Deviation. RMSD is numerically the popularly known 'standard deviation'.

Now, your students will know how to calculate the mean and standard deviation. Do not introduce the concepts of measures of central tendency and the measures of dispersion in the beginning itself. Go ahead in imparting the working knowledge without explaining the background theories. You have to explain the statistical theories only after the students acquire the working knowledge. Otherwise, the will disappear from your class.

Next, introduce the idea of relationship. Ask: When will you say that two things are related? When will you say that two things are not related? You will get the most hilarious replies. 99.99% of the replies will make you more humorous. Enjoy the nonsensical replies. Now clearly define the condition under which one can say that two things are said to be related.

Two things are said to be related only if the change in one affects the other one. Two things are said to be unrelated if the change in one does not affect the other one.

You and your spouse are said to be related because your activities affect your spouse. If your activities do not affect your spouse, then we cannot say that you two are related.

Stress on the word 'change'. Ask: **Can we identify the relationship between any two things which do not change?** *Ask to give some examples on the statements describing the relations.* Students would tend to make the following types of statements.

1. The concentration is related to O.D

You have to restate the above as: *the change in concentration affects the OD.*

2. Concentration is related to the rate of the reaction

Restate as: *Change in concentration affects the rate of the reaction.*

You have to restate the normal statements of relations in the form: how the change in one affects the other one. Otherwise, your students will not learn the method of identifying the relationship between any two things. The will never understand the role of statistics in establishing the relationship between any two things.

The following relations can be defined, if there exists relationship among the variables.

(mean of x) = **A**. (mean of y) . . . . . (1)

(sd of x) = **B**. (sd of y) . . . . . (2)

Equations 1 and 2 say that what is the change in the mean of x when the mean of y changes; and what is the change in the sd of x when the sd of y changes. Both the changes in the means and in the sd's of x and y has to be taken in to consideration to identify the relationship between x and y. To identify whether there exists any correlation between any two variables, one has to compare the **M**ean to **S**tandard deviation to **R**atios (**MSR**).

From (1), **A** = (mean of x) / (mean of y).

From (2), **B** = (sd of x)/(sd of y).

Correlation coefficient, r = **A** / **B**; Regression coefficient, b = r / **B**

[A/B] = 1, if there exists a perfect positive correction. If there exists a perfect negative correction, [A/B] = -1

Illustrate this with very simple examples

Example: 1

Note: regression coefficient, b, is the slope of the straight line, not the x or y intercept. b is the ratio of difference in y to difference in x.

x	y
1	1
2	2
3	3
4	4
5	5

mean of x = 3; sd of x = 1.58

mean of y = 3; sd of y = 1.58

A = (mean of x) / (mean of y) = 3/3 = 1

B = (sd of x) /(sd of y) = 1.58/1.58 = 1

Correlation Coefficient, r = A/B = 1/1 = 1, which indicates perfect positive correlation.

Regression Coefficient, b = r/B = 1/1 = 1, indicating a graph with positive slope.

Example 2

x y

1 -1

2 -2

3 -3

4 -4

5 -5

mean of x = 3; sd of x = 1.58

mean of y = -3; sd of y = 1.58

A = (mean of x) / (mean of y) = 3/(-3) = (-1)

B = (sd of x) /(sd of y) = 1.58/1.58 = 1

Correlation Coefficient, r = A/B = -(1)/1 =-1, which indicates a perfect negative correlation.

Regression Coefficient, b = r/B =(-1)/1= -1, indicating a graph with negative slope.

Example 3

x y

1 2

2 4

3 6

4 8

5 10

mean of x = 3; sd of x = 1.58

mean of y = 6; sd of y = 3.16

A = (mean of x) / (mean of y) = 3/6 = 0.5

B = (sd of x) /(sd of y) = 1.58/3.16 = 0.5

Correlation Coefficient, r = A/B = 0.5/0.5 = 1, which indicates a perfect positive correlation.

Regression Coefficient, b = r/B =(1)/(0.5)=2, indicating a graph with positive slope.

Example 4

x y

1 -2

2 -4

3 -6

4 -8

5 -10

mean of x = 3; sd of x = 1.58

mean of y = -6; sd of y = 3.16

A = (mean of x) / (mean of y) = 3/(-6) = -0.5

B = (sd of x) /(sd of y) = 1.58/3.16 = 0.5

Correlation Coefficient, r = A/B= -0.5/0.5 =-1, which indicates a perfect negative correlation.

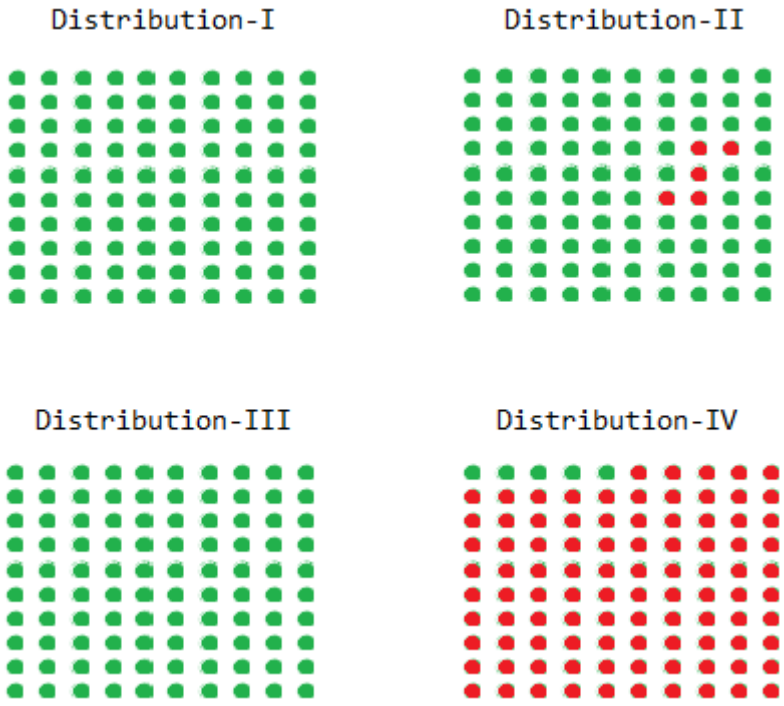
Regression Coefficient, b = r/B=(-1)/(0.5)=-2, indicating a graph with negative slope.

After illustrating all the preceding examples, now ask about their body weights and the amount of money they have. Ask them to consider x as body weight and y as the amout of money they have. Ask them to find whether any relationship exist between their body weights and the amount of money they have. Probably, the correlation value near to zero. Now you explain that there exists no relationship between the variables if the correlation coefficient becomes zero.

Now you ask about their body weights and heights. Tabulate them in the board. Ask them to find whether any correlation exist between the body weight and their height. Give as many examples as possible till they master the art of calculating mean, standard deviation, correlation coefficient and regression coefficient. **Do not start statistical theory till your students become masters in computing the mean, standard deviation, correlation coefficient, and regression coefficient.** Before starting the theoretical statistics, introduce how to find whether the difference between any two means is significant or not. Here 'any two mean' means:

1. population mean vs. sample mean,

In order to make your students understand the significance of difference between the means of any two distributions, explain the following diagram very clearly. Otherwise, they will not understand the 'significance of difference' in the first place.



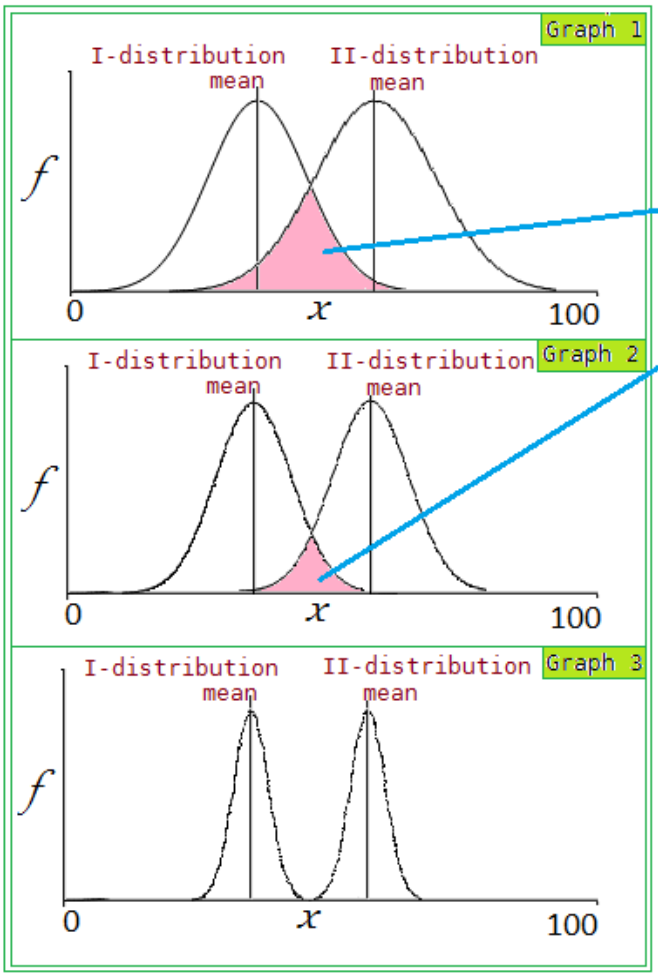
The total number of dots in each distribution is 100. 5 points in distribution-II differs from the points in distribution-I. The size of the difference between distributions I and II is 5/100. 95 points in distribution-IV differs from distribution-III. The size of difference between distributions III and IV is 95/100. The size of difference between distributions III and IV is greater than the size of the difference between distributions I and II. The size of difference is directly proportional to the number of points in which the distributions differ.

The size of difference will be more if the number of points in which the distributions differ is more. The size of difference will be less if the number of points in which the distributions differ is less.

We say that the difference between distributions III and IV is relatively more significant than that of distributions I and II because there are more points of difference between distributions III and IV as compared to the number of points of difference between distributions I and II.

Do not make any attempt to explain the significance of difference between the means of distributions if your students fail to understand the above point. If you are successful, go on to explain the significance of difference between the means of distributions.

Significance of difference between the means of any two distributions



Distributions do not differ from each other at the points in the overlapping area.

Lesser will be the number of points at which the distributions differ, if they more area of overlap. More will be number of points at which the distributions differ, if they have a lesser area of overlap.

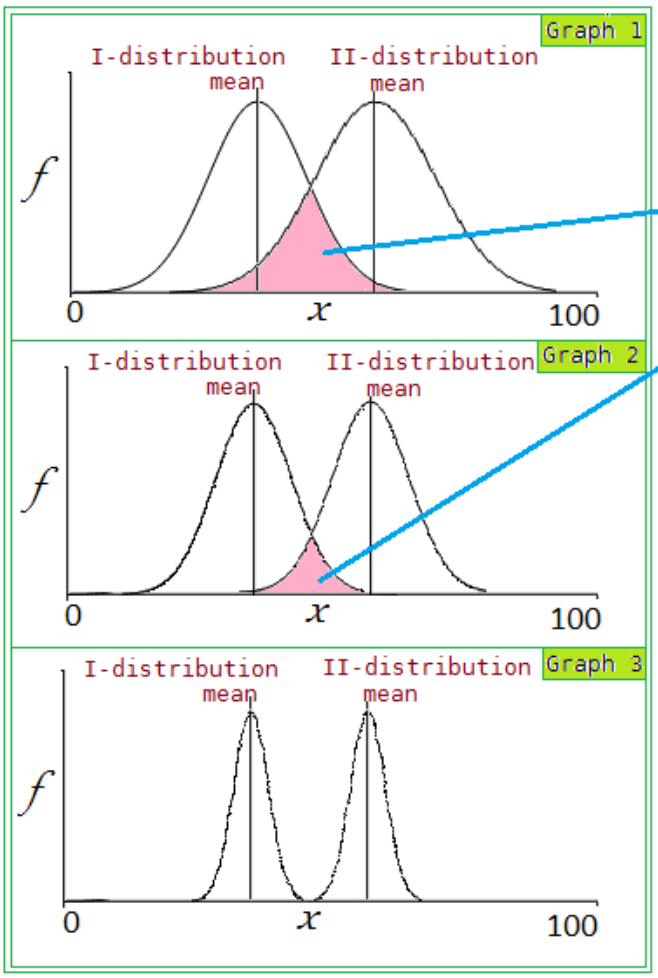
More will be significance of difference between the means of distributions if they differ at more number of points.

Examine the details of the above three graphs. You will find that in each graph there are two distributions. The **difference between the means** of the distributions is same in these graphs. What about the **significance of the difference** between the means in above three graphs? Are they equally significant because they have the same value?

In graph 3, the **distributions do not overlap**. Every point in distribution-II differs from distribution-I. In graph 1 and 2, the **distributions overlap**. The distributions differ only at the points where there is no overlapping. The distributions do not differ at the points in the overlapping area. The significance of difference between the distributions which have less overlapping area will be more because they will have more number of points of difference. The significance of difference between the distributions which have more overlapping area will be less because they will have lesser number of points of difference.

The significance of the difference between the means of any two distributions can be quantitatively assed by taking the ratio of (the difference between the means) to (the standard deviation of difference in the means of distributions).

Significance of difference between the means of sample and population



Distributions do not differ from each other at the points in the overlapping area.

Lesser will be the number of points at which the distributions differ, if they more area of overlap. More will be number of points at which the distributions differ, if they have a lesser area of overlap.

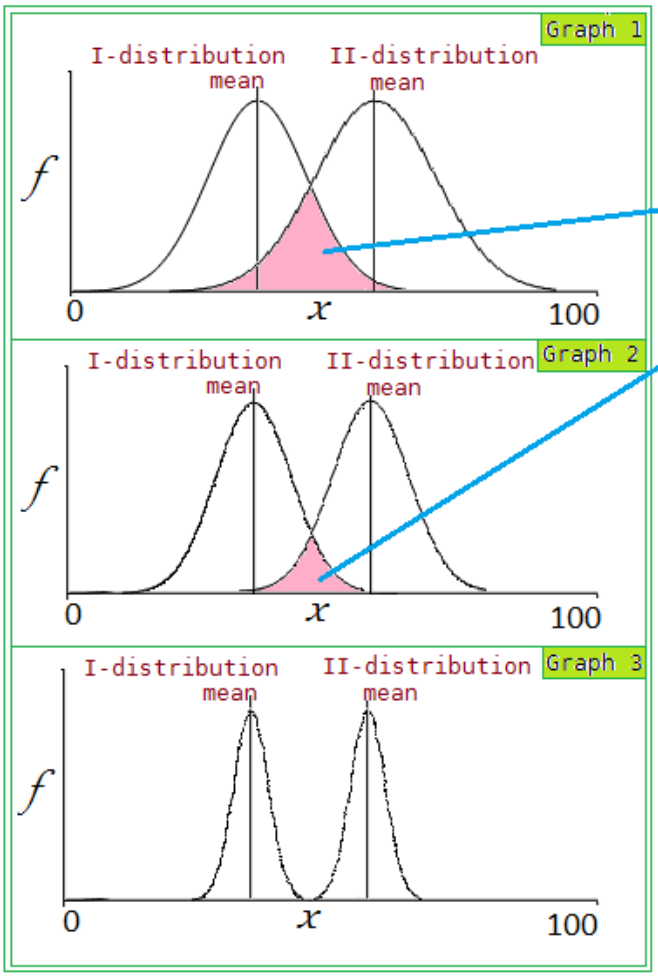
More will be significance of difference between the means of distributions if they differ at more number of points.

Let distribution-I is the values of the sample and distribution-II is the values of the population.

If the sample is a true representative of the population, then the difference in their means has to be insignificant. That means, the distributions should have more area of over-lap.

If the sample is not a true representative of the population, then the difference in their means will be significant. That means, the distributions will have less area of overlap.

Significance of difference between means of samples from a population



Distributions do not differ from each other at the points in the overlapping area.

Lesser will be the number of points at which the distributions differ, if they more area of overlap. More will be number of points at which the distributions differ, if they have a lesser area of overlap.

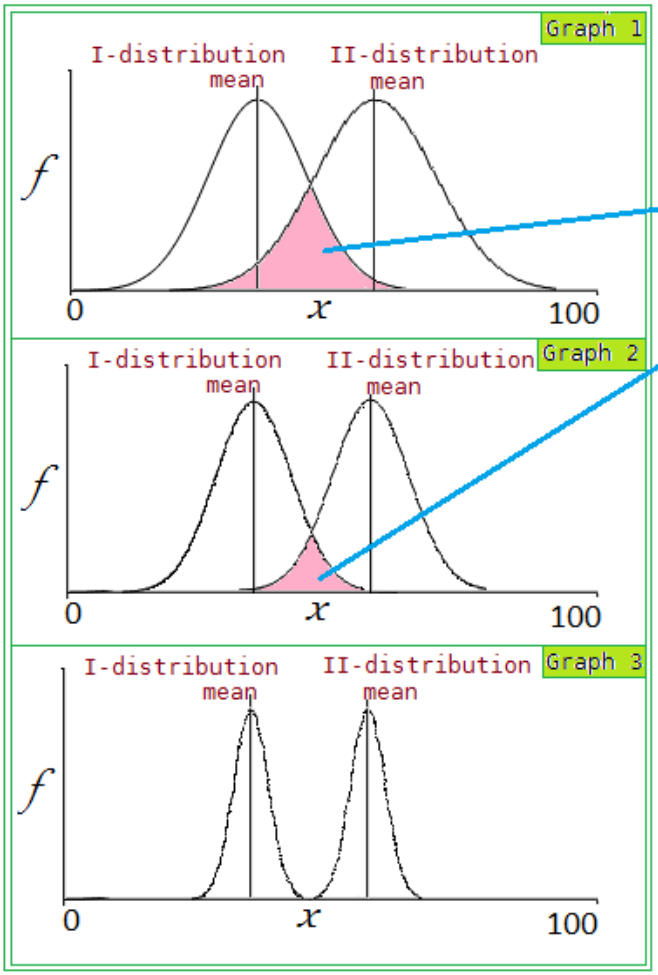
More will be significance of difference between the means of distributions if they differ at more number of points.

Let distribution-I and Distribution represent values of the samples from the same population.

If the samples are from the same population, then the difference in their means has to be insignificant. That means, the distributions should have more area of overlap.

If there is error in sampling, then the difference in their mean will be significant. The distributions will have lesser area of overlap.

Significance of difference between means of Control and Sample



Distributions do not differ from each other at the points in the overlapping area.

Lesser will be the number of points at which the distributions differ, if they more area of overlap. More will be number of points at which the distributions differ, if they have a lesser area of overlap.

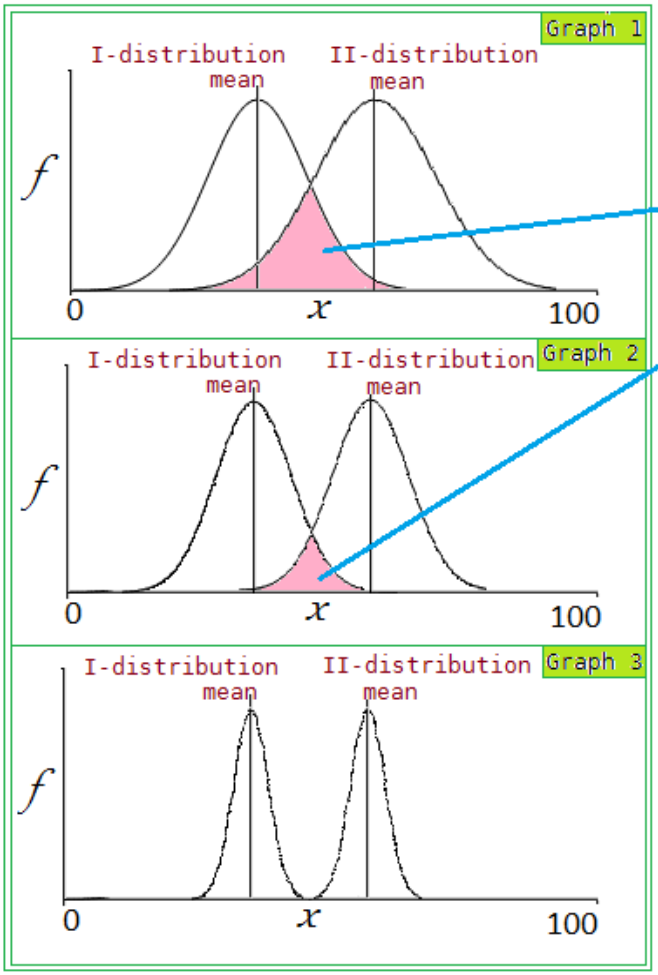
More will be significance of difference between the means of distributions if they differ at more number of points.

Let distribution-I is the values of the from control and distribution-II is the values of the sample.

If the samples differ from control, then the difference in their means has to be significant. That means, the distributions should have less area of overlap.

If the samples are not different from the control, then the difference in their means will be insignificant. That means, the distributions will have more area of overlap.

Significance of difference between means of a sample before and after a treatment



Distributions do not differ from each other at the points in the overlapping area.

Lesser will be the number of points at which the distributions differ, if they more area of overlap. More will be number of points at which the distributions differ, if they have a lesser area of overlap.

More will be significance of difference between the means of distributions if they differ at more number of points.

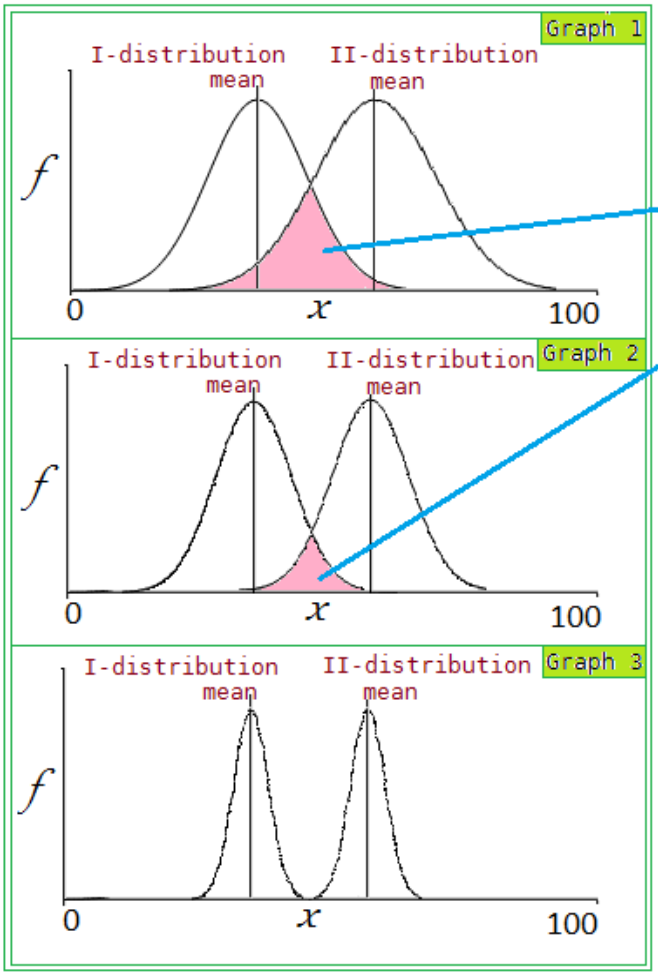
Let distribution-I is the values of a sample before treatment and distribution-II is the values of the sample after a treatment.

If there is a difference sample before and after treatment, then the difference in their means has to be significant. That means, the distributions should less area of overlap.

If the sample is not affected by the treatment, then the difference in their means will be insignificant. That means, the distributions will have more area of overlap.

Significance of difference between

Expected and observed mean



Distributions do not differ from each other at the points in the overlapping area.

Lesser will be the number of points at which the distributions differ, if they more area of overlap. More will be number of points at which the distributions differ, if they have a lesser area of overlap.

More will be significance of difference between the means of distributions if they differ at more number of points.

Let distribution-I is the theoretically computed expected values and distribution-II is the values experimentally recorded values.

If there is no difference between theoretically computed values and experimentally recorded values, then the difference in their means has to be insignificant. That means, the distributions should have more area of overlap. Then, the experiment proves the theory.

If the experimentally recorded values will not be not in accordance with the theoretically computed values. Then, the difference in their means will be significant and the distributions will have less area of overlap.

Quantification of Significance of difference between the means

The significance of difference between the means of can be quantified by through several tests like t-test, z-test, chi square test and f-test and other similar tests.

In all the tests, a test statistic is computed and compared to a theoretical distribution.

$$\text{Test statistic} = \frac{\text{Difference between the means of the distributions}}{\text{Standard deviation of difference of means of distributions}}$$

The test statistic observed has to be compared to the theoretically computed expected static.

If the observed test static is greater than the theoretically computed expected statistic, then the difference is significant, otherwise it is insignificant.

After completing these, start teaching the actual statistics.

You can make your students to use statistics!